# 2020 Census DAS Update:
# DAS Research and Metrics Evaluation

Michael Hawes
Matt Spence

CSAC
May 25, 2021

Shape
your future
START HERE >

United States®
Census
2020

# Recent Activity: DAS Tuning for the Redistricting Data

**P.L. 94-171 Redistricting Data Summary File Tuning & Privacy-Accuracy Trade-off Experiments**

- In December through March, the DAS Team conducted over 600 full-scale TDA runs with the complete P.L. 94-171 data product schema.
- Goal: Evaluating resulting accuracy of varying parameters for:
  - Overall setting of PLB
  - Query strategy
  - Allocation of PLB across geographic levels
  - Allocation of PLB across queries
- Worked with subject matter experts in Demographic and Decennial Directorates to evaluate accuracy of experimental runs to inform parameter setting.

# Disclosure Avoidance System Tuning

To ensure fitness for use of the Disclosure Avoidance System (DAS) for the redistricting data product, we have thoroughly tuned the system to the redistricting and Voting Rights Act use cases, as submitted to us by redistricting experts and the Department of Justice.

Key metrics used for this tuning were the accuracy of largest racial group, as a proportion of total population within **on-spine and off-spine geographies with total populations of 500-549 persons.**

Tuning target was that this proportion is **within 5 percentages points of the enumerated value ≥ 95% of the time**.

# Experiment Design
# Dimension 1: Query Strategy

| Query Strategy 1 |
|---|
| TOTAL POPULATION |
| CENRACE |
| HISPANIC |
| VOTINGAGE |
| HHINSTLEVELS |
| HHGQ |
| HISPANIC*CENRACE |
| VOTINGAGE*CENRACE |
| VOTINGAGE*HISPANIC |
| VOTINGAGE*HISPANIC*CENRACE |
| DETAILED (HHGQ x VOTING_AGE x HISPANIC x CENRACE) |

VS.

| Query Strategy 2 |
|---|
| TOTAL POPULATION |
| HISPANIC***CENRACE11CATS** |
| VOTINGAGE |
| HHINSTLEVELS |
| HHGQ |
| HISPANIC***CENRACE11CATS***VOTINGAGE |
| DETAILED (HHGQ x VOTING_AGE x HISPANIC x **CENRACE**) |

# Experiment Design
# Dimension 2:  PLB Allocation

| PLB ALLOCATION A |
|---|
| Equal share of PLB to each query in strategy |

vs.

| PLB ALLOCATION B |
|---|
| Variable share of PLB to each query by query size |

# Experiment Design Dimension 3: Geographic Spine

**AIAN Spine**

Isolates post-processing for AIAN Tribal Areas within each state.

vs.

**Optimized Spine**

Optimized version of the AIAN spine that redefines geographies to approximate off-spine geographic entities of interest, and isolates GQs into their own block groups by type.

# Experiments
# Minimal PLB to Meet Accuracy Targets

| Query Strategy | PLB Allocation | Spine | Minimal PLB to Meet Targets |
|---|---|---|---|
| 1 | A | AIAN Spine | 8.29 |
| 1 | A | Opt-Spine | 7.88 |
| 1 | B | AIAN Spine | 10.30 |
| 1 | B | Opt-Spine | 8.92 |
| 2 | A | AIAN Spine | 7.94 |
| 2 | A | Opt-Spine | 9.33 |
| 2 | B | AIAN Spine | 9.75 |
| 2 | B | Opt-Spine | 9.18 |

# Evaluating the results & Selecting a Strategy

Reviewed a range of metrics at various geographic levels

- Total Population
- Total Population Aged 18+
- Race Alone
- Race Alone or In Combination
- Hispanic/Not Hispanic
- Hispanic*Race Alone
- Group Quarters (GQ) Type

Also considered implications of the strategies' accuracy across a range of metrics for the extension of P.L. data to the fuller Demographic and Housing Characteristics files.

Selected Query Strategy 1 (full CENRACE variable), PLB Allocation B (proportional by query size), with the Optimized Geographic Spine.

Allocated additional PLB to the Block Group level.

# Demonstration Data

- Since October 2019, the Census Bureau has been periodically releasing demonstration data products (using 2010 Census data) for data user evaluation.
- The first four of these sets of demonstration data (October 2019, May 2020, September 2020, November 2020) used a conservative global PLB set by DSEP for the October 2019 Demonstration Product, in order to evaluate algorithmic improvements.
- ***The 2020 Census Data Products will not be held to this fixed PLB.***

- On April 28, 2021 we released another set of Privacy-Protected Microdata Files (PPMFs) and Detailed Summary Metrics using a different global PLB (ε=12.2) that more closely approximates the level of PLB that the DSEP will be considering for the 2020 Census redistricting data files.
- Exceeded the established accuracy targets: for places and other off-spine entities with populations between 500-549 people, 99.52% of these geographies meet the accuracy target; those with larger populations performed even better.
- In September, we plan to release a final set of PPMFs using the actual production code and settings that will be used for the 2020 Census redistricting data files.

# April 2021 PPMF Privacy-loss Budget Allocation (by geographic level)

| Privacy-loss Budget Allocation April 28, 2021<br>PPMF<br>Person Tables (PPMF-P)<br>United States | | |
|---|---|---|
| | | |
| | Global rho | 192721/184041 (1.05) |
| | Global epsilon | 10.3 |
| | delta | $10^{-10}$ |
| | | |
| | | rho Allocation by Geographic Level |
| | US | 51/1024 |
| | State | 153/1024 |
| | County | 78/1024 |
| | Tract | 51/1024 |
| | Optimized block group* | 172/1024 |
| | Block | 519/1024 |

| Privacy-loss Budget Allocation April 28, 2021<br>PPMF<br>Units Tables (PPMF-U)<br>United States | | |
|---|---|---|
| | | |
| | Global rho | 919681/20241001 (0.045) |
| | Global epsilon | 1.9 |
| | delta | $10^{-10}$ |
| | | |
| | | rho Allocation by Geographic Level |
| | US | 1/1024 |
| | State | 1/1024 |
| | County | 18/1024 |
| | Tract | 75/1024 |
| | Optimized block group* | 906/1024 |
| | Block | 23/1024 |

*Optimized block groups do not affect tabulation geography.

# April 2021 PPMF Privacy-loss Budget Allocation (by query)

| Query | Per Query rho Allocation by Geographic Level | | | | | |
|---|---|---|---|---|---|---|
| | US | State | County | Tract | Optimized Block Group* | Block |
| TOTAL (1 cell) | | 678/1024** | 342/1024 | 1/1024 | 572/1024 | 1/1024 |
| CENRACE (63 cells) | 2/1024 | 1/1024 | 1/1024 | 2/1024 | 1/1024 | 2/1024 |
| HISPANIC (2 cells) | 1/1024 | 1/1024 | 1/1024 | 1/1024 | 1/1024 | 1/1024 |
| VOTINGAGE (2 cells) | 1/1024 | 1/1024 | 1/1024 | 1/1024 | 1/1024 | 1/1024 |
| HHINSTLEVELS (3 cells) | 1/1024 | 1/1024 | 1/1024 | 1/1024 | 1/1024 | 1/1024 |
| HHGQ (8 cells) | 1/1024 | 1/1024 | 1/1024 | 1/1024 | 1/1024 | 1/1024 |
| HISPANIC*CENRACE (126 cells) | 5/1024 | 2/1024 | 3/1024 | 5/1024 | 3/1024 | 5/1024 |
| VOTINGAGE*CENRACE (126 cells) | 5/1024 | 2/1024 | 3/1024 | 5/1024 | 3/1024 | 5/1024 |
| VOTINGAGE*HISPANIC (4 cells) | 1/1024 | 1/1024 | 1/1024 | 1/1024 | 1/1024 | 1/1024 |
| VOTINGAGE*HISPANIC*CENRACE (252 cells) | 17/1024 | 6/1024 | 11/1024 | 17/1024 | 8/1024 | 17/1024 |
| HHGQ*VOTINGAGE* HISPANIC*CENRACE (2,016 cells) | 990/1024 | 330/1024 | 659/1024 | 989/1024 | 432/1024 | 989/1024 |

*The optimized block groups used within the TopDown Algorithm differ from tabulation block groups. These differences improve accuracy for "off-spine" geographies like places and minor civil divisions. The use of optimized block groups for measurement and post-processing within the TopDown Algorithm does not impact how the resulting data will be tabulated. All Census data products will be tabulated using the official tabulation block groups as defined by the Census Bureau's Geography Division.

**The TOTAL query (total population) is held invariant at the state level. This rho allocation assigned to TOTAL at the state level is the amount assigned to the state-level queries for the total population of all American Indian and Alaska Native (AIAN) tribal areas within the state and for the total population of the remainder of the state, for the 36 states that include AIAN tribal areas.

# Evaluating DAS Runs using Detailed Summary Metrics

Matthew Spence

Population Division

# Detailed Summary Metrics on Demographic Reasonableness

- Compare tabulated quantities at various geographic levels (e.g., Total Population at the county level, Asian Alone at the tract level) for a DAS run to published (i.e., swapped) 2010 tabulations.

- Metrics released with April 2021 PPMF include:

| | |
|---|---|
| **Accuracy** | - Mean Absolute Error (MAE): What is the average absolute change (+/-)?<br>- Mean Absolute Percent Error (MAPE): What is the average relative change (+/- %) |
| **Bias** | - Mean Error (ME): What is the average directional change?<br>- Mean Algebraic Percent Error (MALPE): What is the average directional relative change? |
| **Outliers** | - How many geographies are above particular thresholds? |

# Calculating Metrics: MAE

| County | Published 2010 Population | PPMF 2010 Population | Error | Absolute Error |
|---|---:|---:|---:|---:|
| Autauga County, Alabama | 54,571 | 54,581 | 10 | 10 |
| Baldwin County, Alabama | 182,265 | 182,263 | -2 | 2 |
| Barbour County, Alabama | 27,457 | 27,455 | -2 | 2 |
| Bibb County, Alabama | 22,915 | 22,922 | 7 | 7 |
| Blount County, Alabama | 57,322 | 57,321 | -1 | 1 |
| … | … | … | … | … |
| Loving County, Texas | 82 | 77 | -5 | 5 |
| … | … | … | … | … |

Mean Absolute Error: 4.91

# Calculating Metrics: MAPE

| County | Published 2010 Population | PPMF 2010 Population | Percent Error | Absolute Percent Error |
|---|---:|---:|---:|---:|
| Autauga County, Alabama | 54,571 | 54,581 | 0.0183% | 0.0183% |
| Baldwin County, Alabama | 182,265 | 182,263 | -0.0011% | 0.0011% |
| Barbour County, Alabama | 27,457 | 27,455 | -0.0073% | 0.0073% |
| Bibb County, Alabama | 22,915 | 22,922 | 0.0305% | 0.0305% |
| Blount County, Alabama | 57,322 | 57,321 | -0.0017% | 0.0017% |
| … | … | … | … | … |
| Loving County, Texas | 82 | 77 | -6.0976% | 6.0976% |
| … | … | … | … | … |

Mean Absolute Percent Error: 0.04%

# Outliers: Keep Size in Mind

| | Count of Units (N) | Count where the absolute percent difference exceeds 10% | % Over Threshold |
|---|---|---|---|
| All counties | | | |
| White alone | 3,143 | 1 | 0.03% |
| Black alone | 3,143 | 743 | 23.64% |
| AIAN alone | 3,143 | 919 | 29.24% |
| Asian alone | 3,143 | 1,019 | 32.42% |
| NHPI alone | 3,143 | 2,131 | 67.80% |
| SOR alone | 3,143 | 727 | 23.13% |
| Two or more races | 3,143 | 738 | 23.48% |

CBDRB-FY21-POP001-0100.

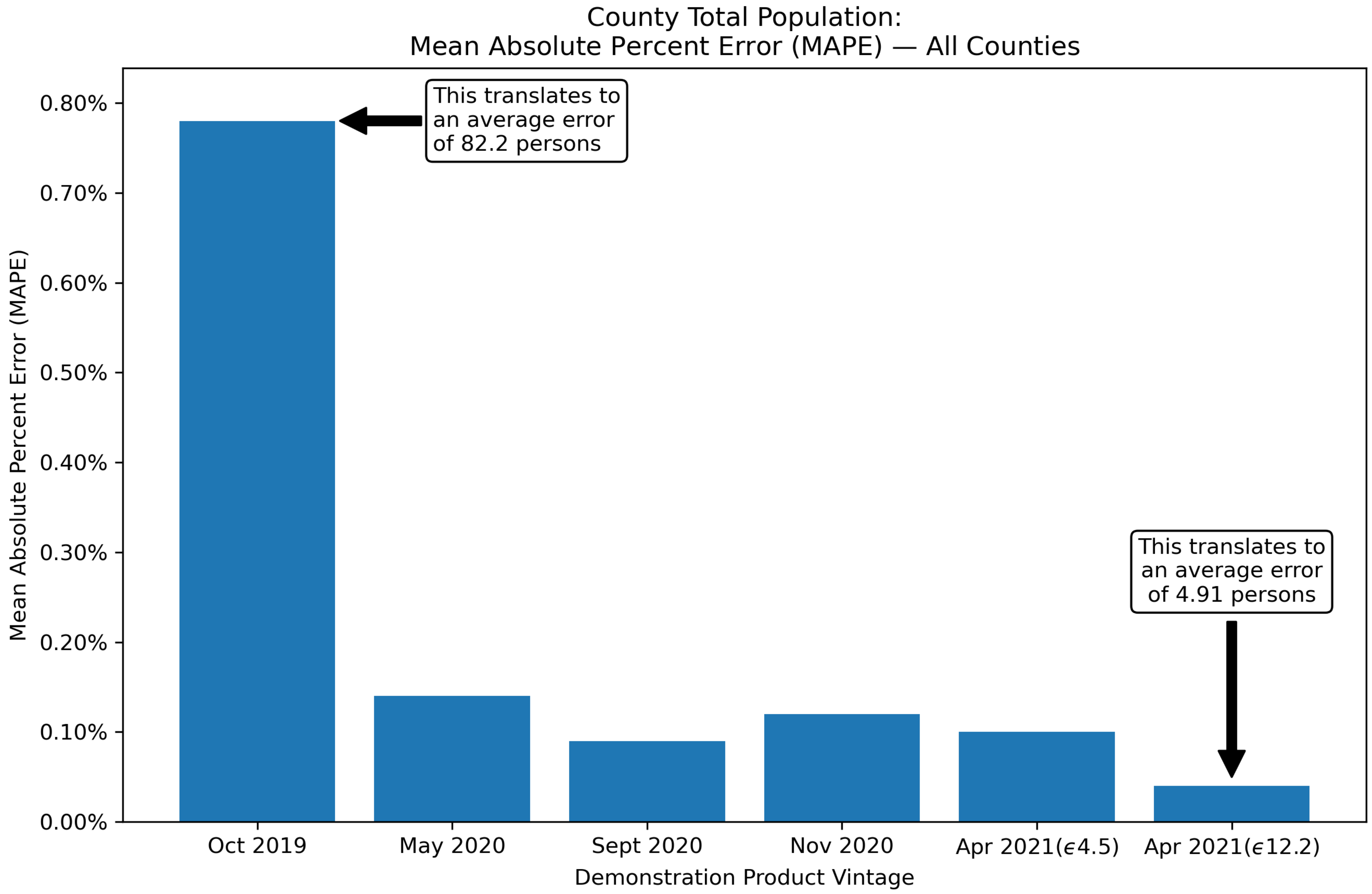
# Outliers: Keep Size in Mind

| | Count of Units (N) | Count where the absolute percent difference exceeds 10% | % Over Threshold | Average Population |
|---|---:|---:|---:|---:|
| All counties | | | | |
| White alone | 3,143 | 1 | 0.03% | 77,127 |
| Black alone | 3,143 | 743 | 23.64% | 12,386 |
| AIAN alone | 3,143 | 919 | 29.24% | 933 |
| Asian alone | 3,143 | 1,019 | 32.42% | 4,669 |
| NHPI alone | 3,143 | 2,131 | 67.80% | 172 |
| SOR alone | 3,143 | 727 | 23.13% | 6,079 |
| Two or more races | 3,143 | 738 | 23.48% | 2,866 |

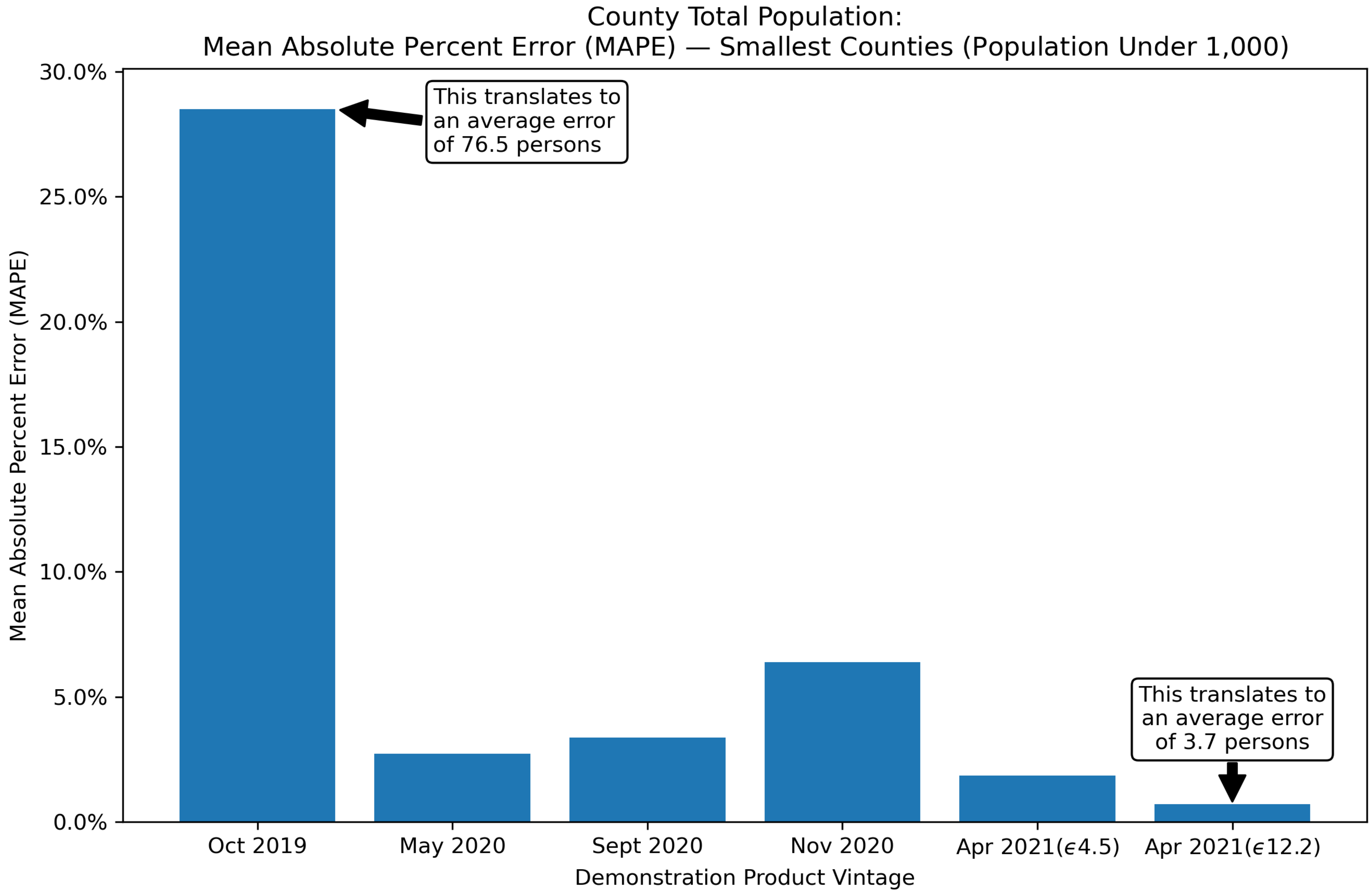
# Outliers: Compare Like with Like

| | Count of Units (N) | Count where the absolute percent difference exceeds 10% | % Over Threshold |
|---|---:|---:|---:|
| Incorporated places **with population 0 to 9** | | | |
| White alone | 99 | 89 | 89.90% |
| Black alone | 9,335 | 6,031 | 64.61% |
| AIAN alone | 11,981 | 8,195 | 68.40% |
| Asian alone | 12,000 | 7,426 | 61.88% |
| NHPI alone | 17,453 | 6,268 | 35.91% |
| SOR alone | 9,989 | 6,643 | 66.50% |
| Two or more races | 7,547 | 6,363 | 84.31% |

# Improvements Since October 2019

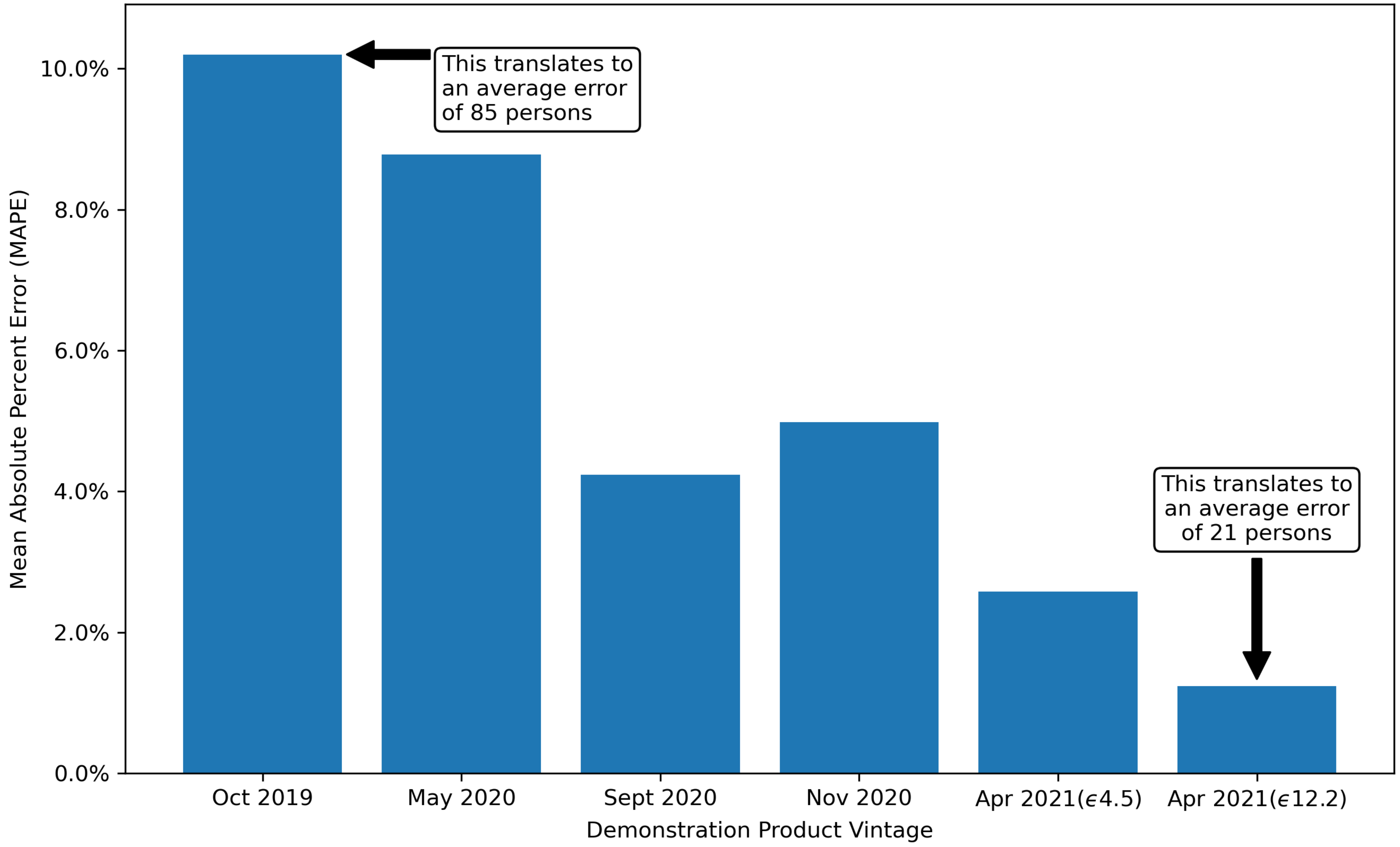


CBDRB-FY20-DSEP-001.



CBDRB-FY20-DSEP-001.

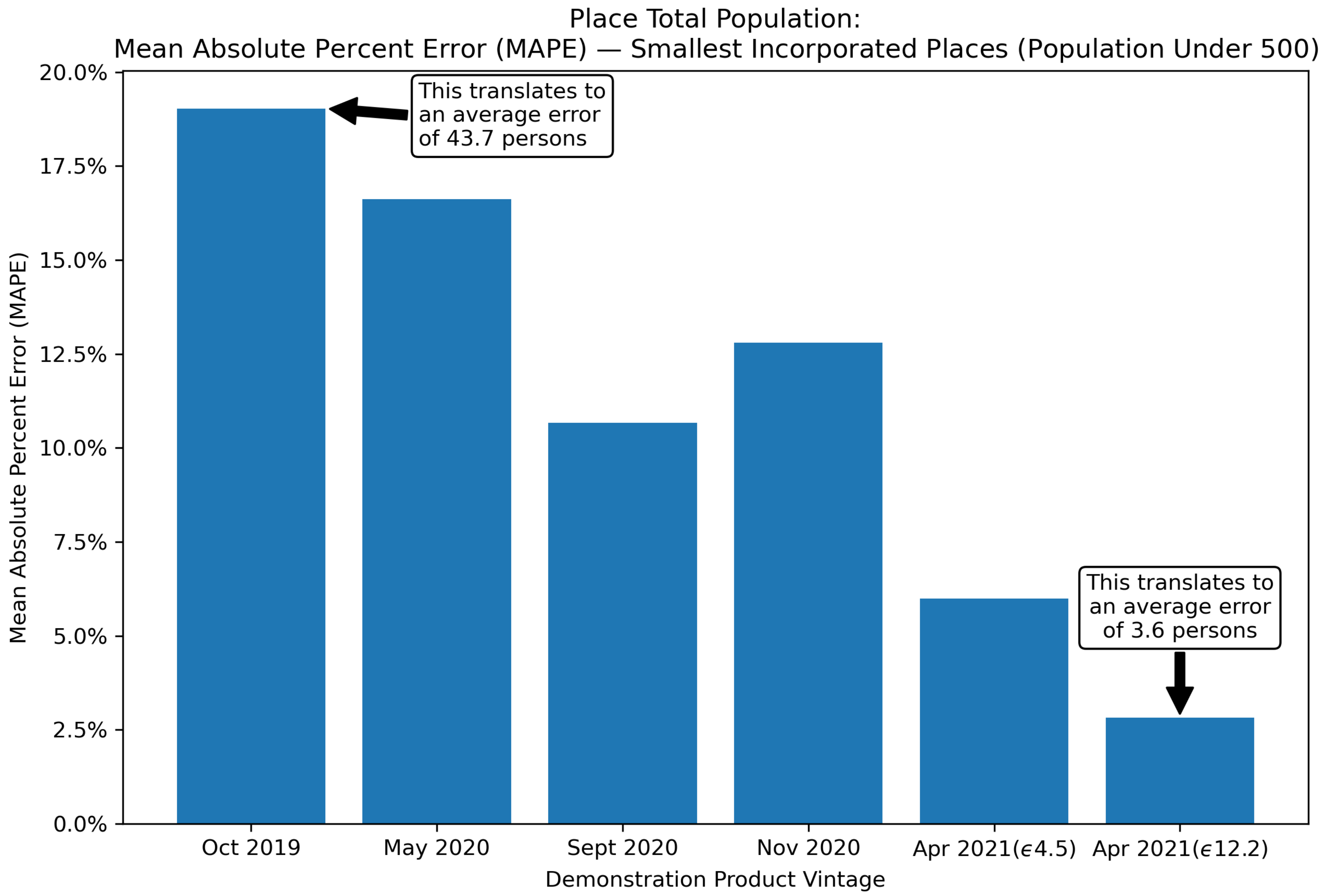
## Place Total Population: Mean Absolute Percent Error (MAPE) — All Incorporated Places



This translates to an average error of 85 persons

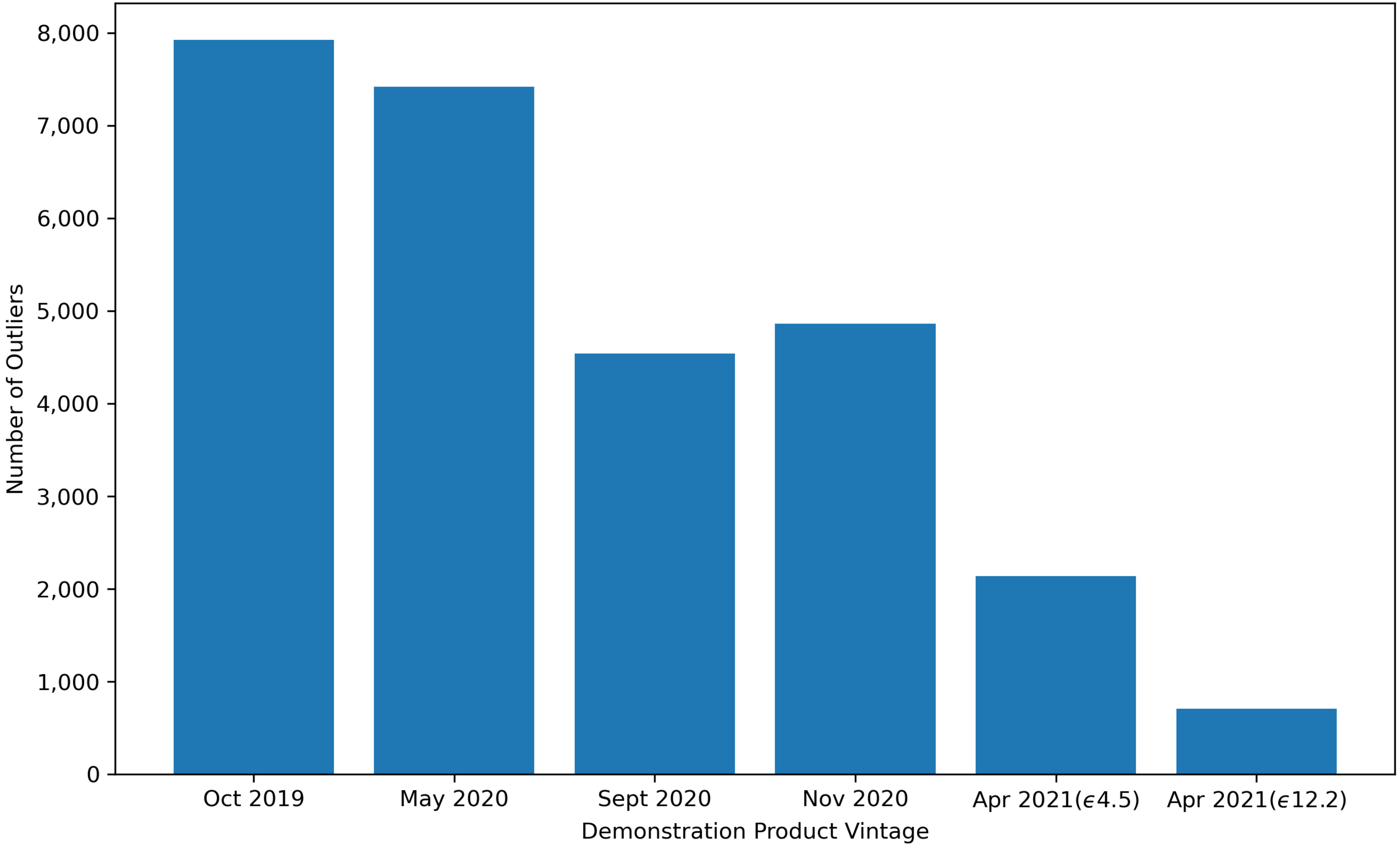
This translates to an average error of 21 persons

CBDRB-FY20-DSEP-001.



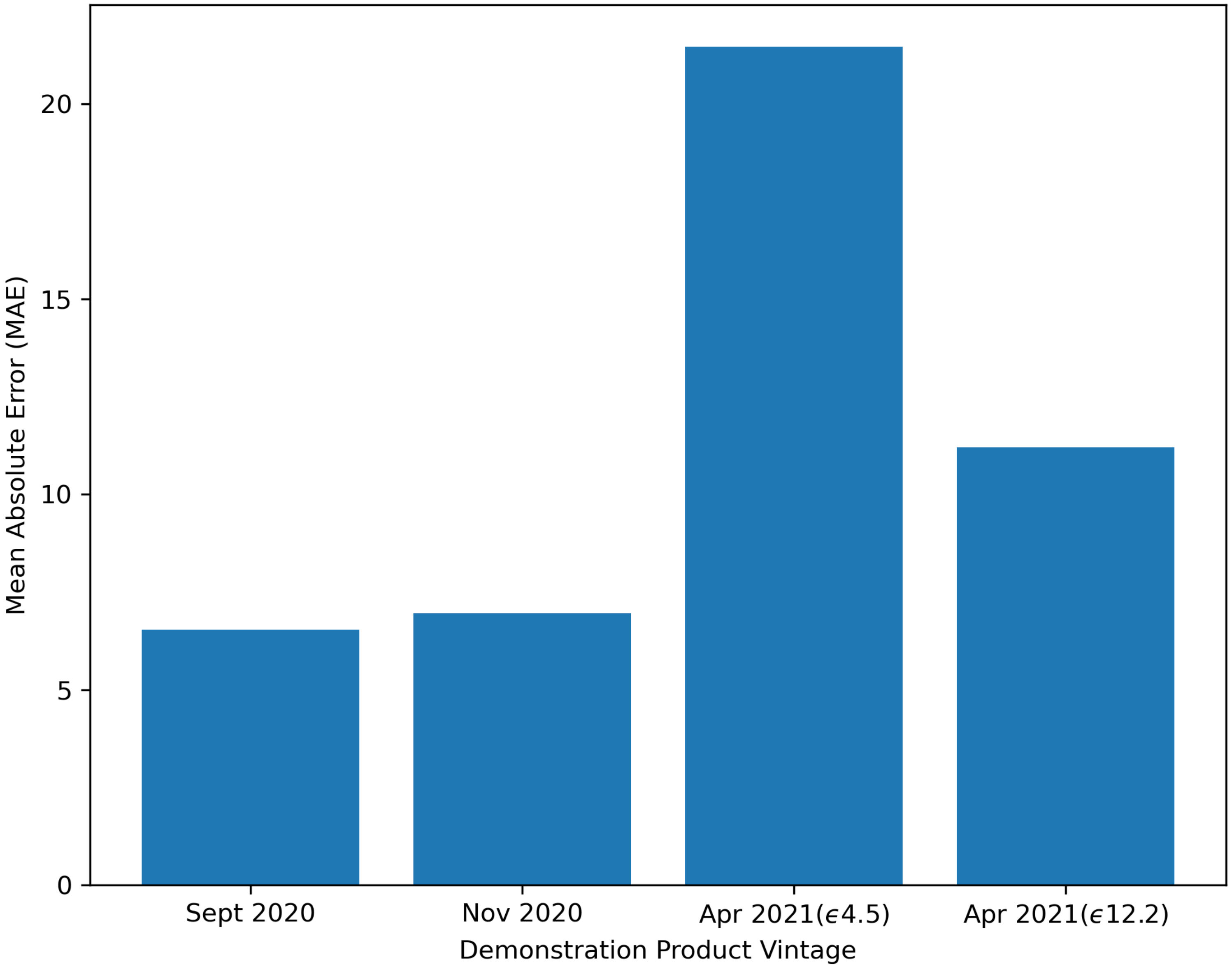
CBDRB-FY20-DSEP-001.

## Place Total Population: Number Exceeding 5% Error — All Incorporated Places

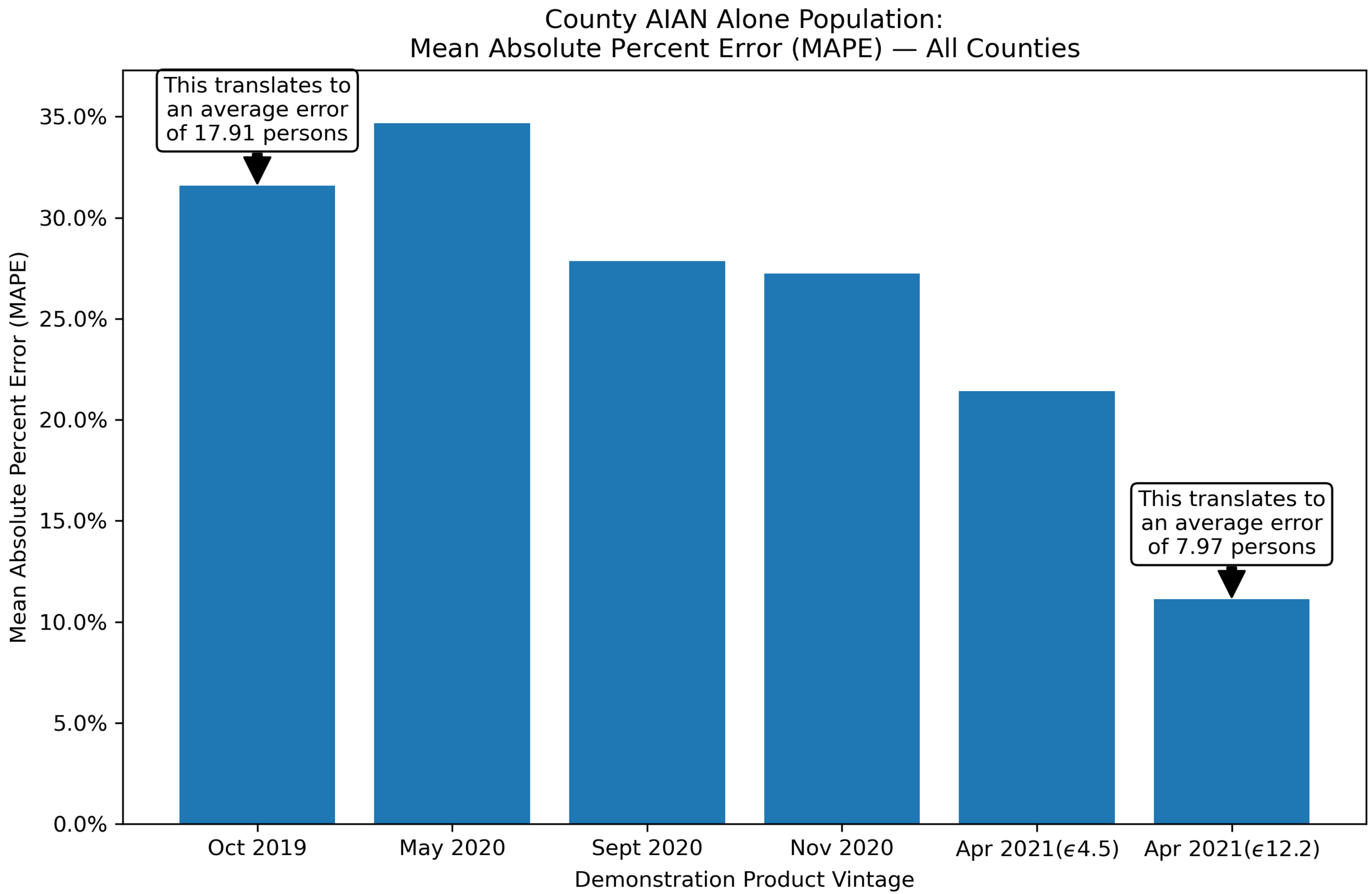


CBDRB-FY20-DSEP-001.

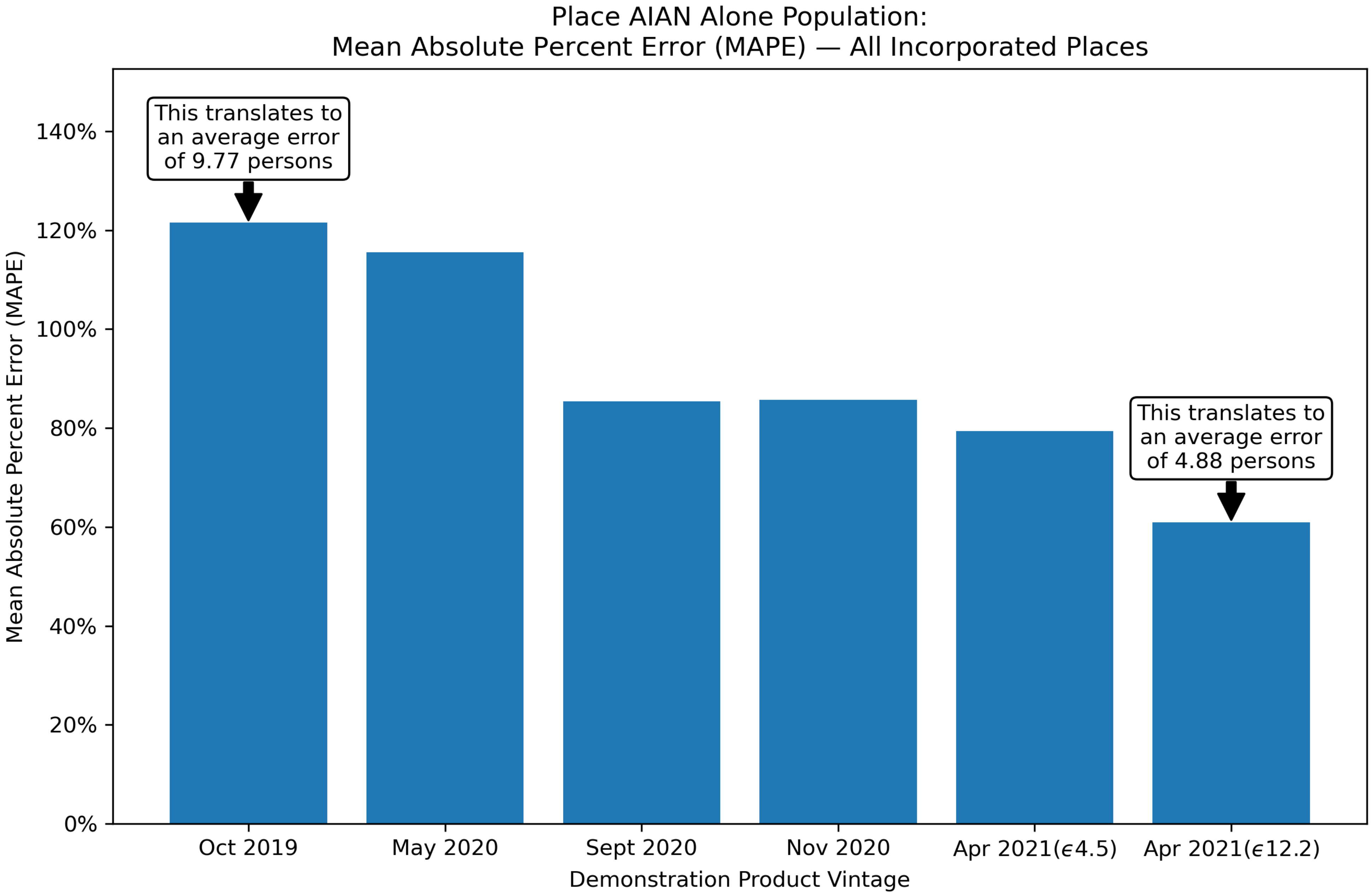
Federal American Indian Reservation/Off-Reservation Trust Land
Total Population: Mean Absolute Error (MAE) — All Areas



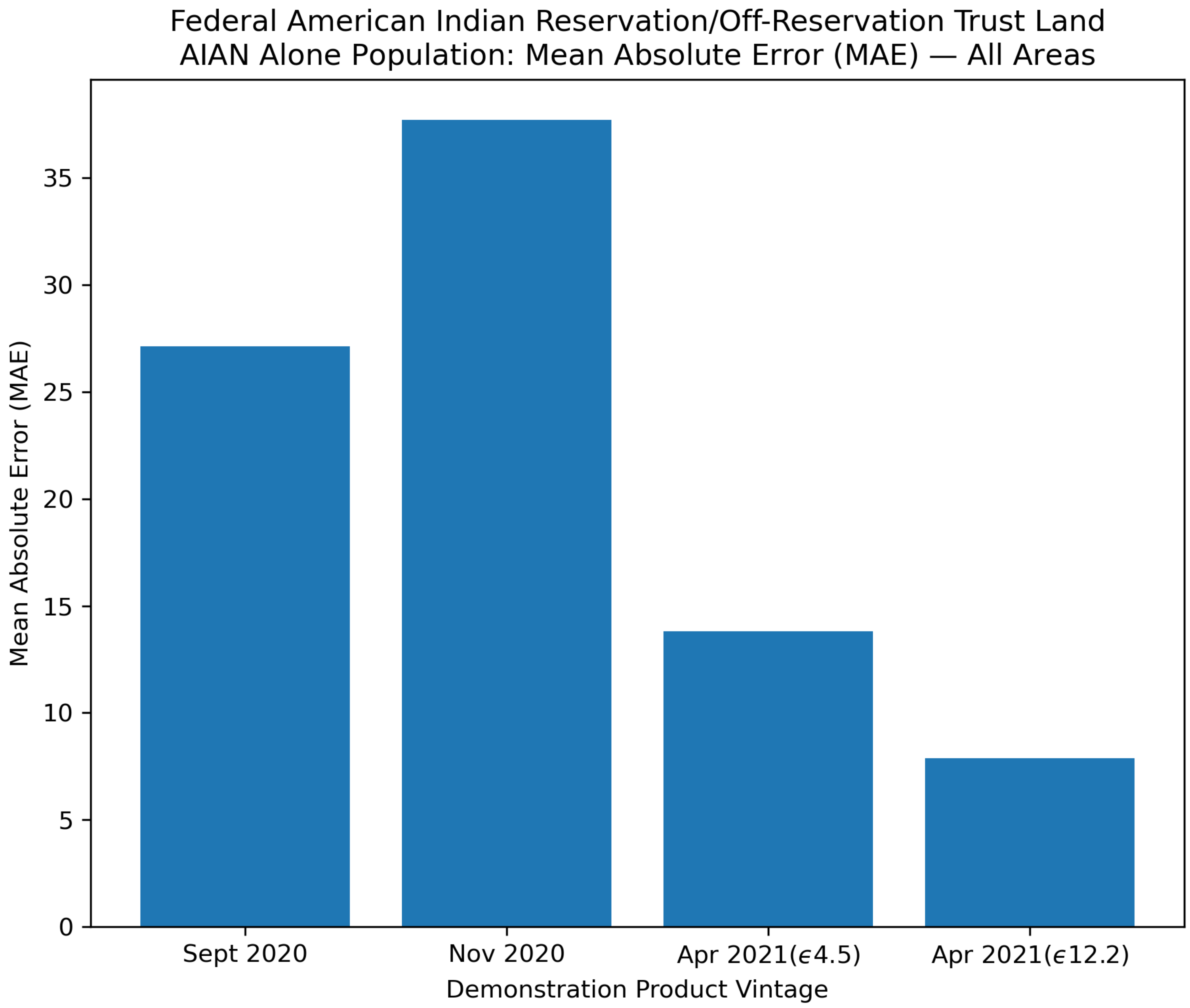
CBDRB-FY20-DSEP-001.



CBDRB-FY20-DSEP-001.



CBDRB-FY20-DSEP-001.



CBDRB-FY20-DSEP-001.

# Where Have Metrics Gotten Worse?

| Metric | Nov 2020 PPMF MAE | April 2021 ϵ 12.2 PPMF MAE |
|---|---:|---:|
| Total Population: Tract | 5.78 | 22.18 |
| Total Population: Puerto Rico Tract | 4.71 | 14.45 |
| Total Population: Elementary School District (ESD) | 24.93 | 37.69 |
| Total Population: Secondary School District (SSD) | 29.69 | 86.48 |
| Total Population: Unified School District (USD) | 32.73 | 48.72 |
| Total Population: Minor Civil Division (MCD) | 17.92 | 18.02 |
| Total Population: Federal American Indian Reservations/Off-Reservation Trust Land (Fed AIR) | 6.95 | 11.20 |
| Total Population: Oklahoma Tribal Statistical Area | 46.79 | 50.86 |
| Two or More Races: State | 20.71 | 44.43 |
| Two or More Races: Incorporated Place | 16.40 | 17.67 |
| Two or More Races: Tract | 14.03 | 16.02 |

CBDRB-FY20-DSEP-001.

# Other Areas of Concern

- Occupied/Vacant Statistics from Unit file still seem noisy relative to $\epsilon$
- Inconsistencies Between MDF Unit and MDF Person Files:
  - 10.8% of blocks with at least one household person have zero occupied housing units
  - 3.6% of blocks with at least one occupied housing unit have more occupied housing units than household persons.
- Improbable Results:
  - 1.5% of blocks (all with no GQ population) feature everyone aged 17 and younger
  - 1.5% of blocks have >10 Persons Per Household (PPH)

CBDRB-FY20-DSEP-001.

# How to Submit Feedback

The changes in the April 2021 PPMFs data set reflect the cumulative feedback received from the data user community throughout the development process. We look forward to feedback from data users on this new demonstration product. Your input will inform the Census Bureau’s June 2021 final decision on the PLB and on the 2020 Census redistricting data parameters. **The deadline to submit feedback is May 28, 2021.**

**** Please send comments to 2020DAS@census.gov with the subject line “April 2021 Demonstration Data.”**

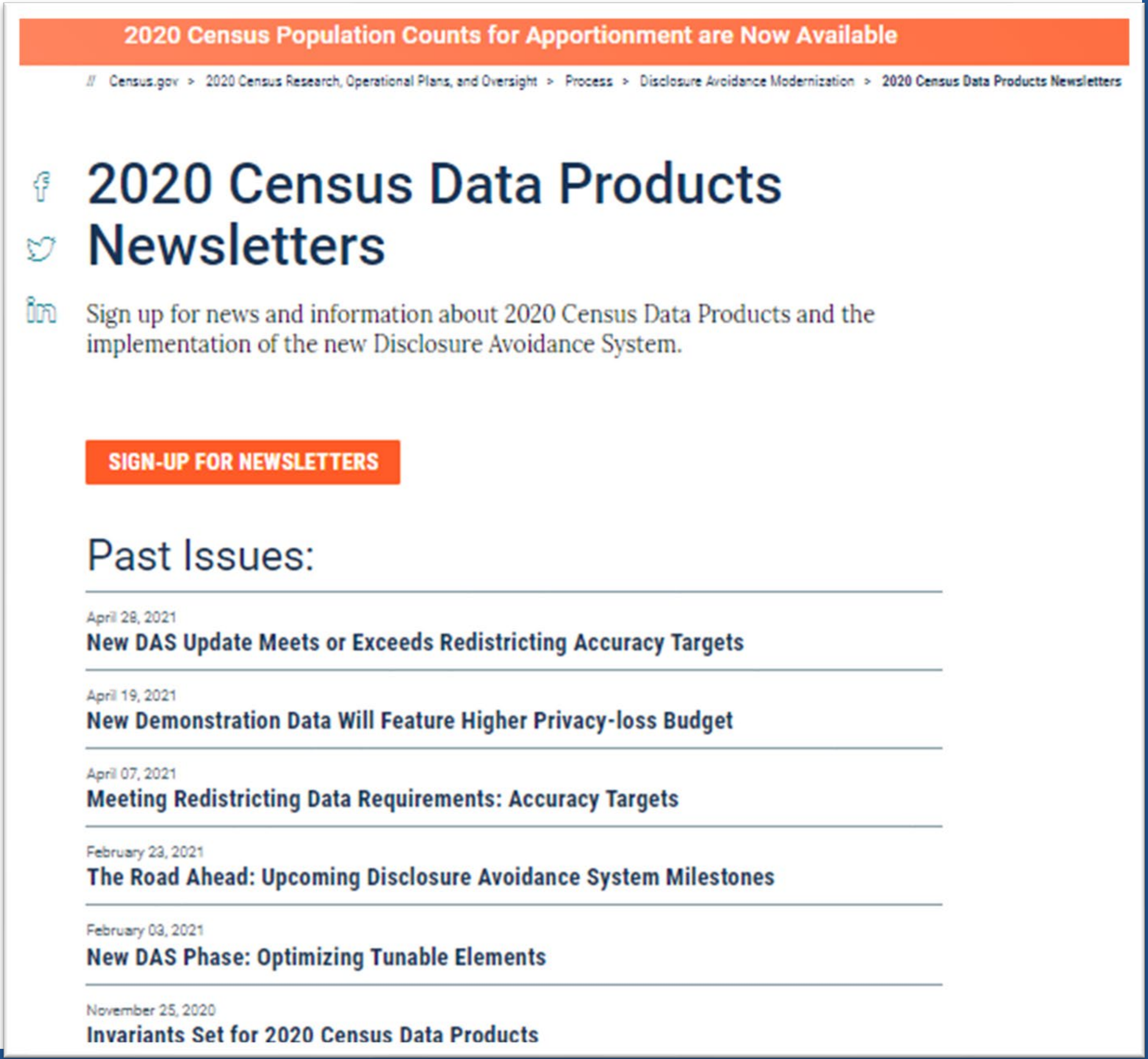
Particularly useful feedback would describe:

- **Fitness-for-use:** Based on your analysis, would the data needed for your applications (redistricting, Voting Rights Act analysis, estimates, projections, funding data sets, etc.) be satisfactory?
    - How did you come to that conclusion?
    - If your analysis found the data to be unsatisfactory, how incrementally would accuracy need to change to improve the use of the data for your required or programmatic use case(s)?
    - Have you identified any improbable results in the data that would be helpful for us to understand?"
- **Privacy:** Do the proposed products present any confidentiality concerns that we should address in the DAS?
- **Improvements:** Are there improvements you’ve identified that you want to make sure we retain in the final design? Be specific about the geography and error metric for the proposed improvement.

Stay Informed:

## Subscribe to the 2020 Census Data Products Newsletters

*Search “Disclosure Avoidance” at www.census.gov



2020 Census Population Counts for Apportionment are Now Available

// Census.gov > 2020 Census Research, Operational Plans, and Oversight > Process > Disclosure Avoidance Modernization > 2020 Census Data Products Newsletters

# 2020 Census Data Products Newsletters

Sign up for news and information about 2020 Census Data Products and the implementation of the new Disclosure Avoidance System.

SIGN-UP FOR NEWSLETTERS

## Past Issues:

April 28, 2021
New DAS Update Meets or Exceeds Redistricting Accuracy Targets

April 19, 2021
New Demonstration Data Will Feature Higher Privacy-loss Budget

April 07, 2021
Meeting Redistricting Data Requirements: Accuracy Targets

February 23, 2021
The Road Ahead: Upcoming Disclosure Avoidance System Milestones

February 03, 2021
New DAS Phase: Optimizing Tunable Elements

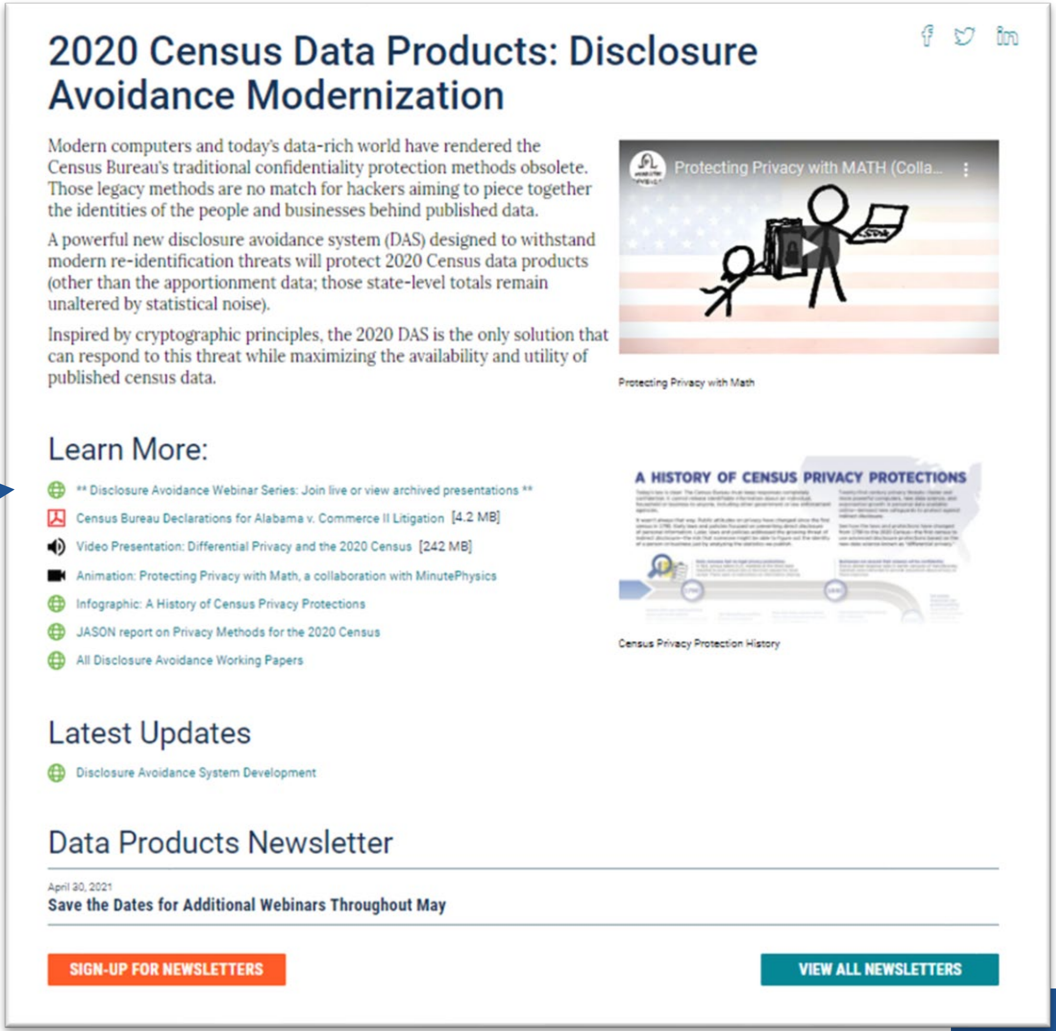
November 25, 2020
Invariants Set for 2020 Census Data Products

Stay Informed:

# Visit Our Website

*Search “Disclosure Avoidance” at www.census.gov

***“Disclosure Avoidance Webinar Series: view archived presentations”***



## 2020 Census Data Products: Disclosure Avoidance Modernization

Modern computers and today's data-rich world have rendered the Census Bureau's traditional confidentiality protection methods obsolete. Those legacy methods are no match for hackers aiming to piece together the identities of the people and businesses behind published data.

A powerful new disclosure avoidance system (DAS) designed to withstand modern re-identification threats will protect 2020 Census data products (other than the apportionment data; those state-level totals remain unaltered by statistical noise).

Inspired by cryptographic principles, the 2020 DAS is the only solution that can respond to this threat while maximizing the availability and utility of published census data.

Protecting Privacy with Math

### Learn More:

- ** Disclosure Avoidance Webinar Series: Join live or view archived presentations **
- Census Bureau Declarations for Alabama v. Commerce II Litigation [4.2 MB]
- Video Presentation: Differential Privacy and the 2020 Census [242 MB]
- Animation: Protecting Privacy with Math, a collaboration with MinutePhysics
- Infographic: A History of Census Privacy Protections
- JASON report on Privacy Methods for the 2020 Census
- All Disclosure Avoidance Working Papers

Census Privacy Protection History

### Latest Updates

- Disclosure Avoidance System Development

### Data Products Newsletter

April 30, 2021

**Save the Dates for Additional Webinars Throughout May**

# Questions?

Shape
your future
START HERE >

United States Census 2020